Disentangling Sources of Gene Tree Discordance in Phylogenomic Datasets: Testing Ancient Hybridizations in Amaranthaceae s.l.

Diego F. Morales-Briones[1*], Gudrun Kadereit[2], Delphine T. Tefarikis[2], Michael J. Moore[3], Stephen A. Smith[4], Samuel F. Brockington[5], Alfonso Timoneda[5], Won C. Yim[6], John C. Cushman[6], Ya Yang[1*]

[1] Department of Plant and Microbial Biology, University of Minnesota-Twin Cities, 1445 Gortner Avenue, St. Paul, MN 55108, USA

[2] Institut für Molekulare Physiologie, Johannes Gutenberg-Universität Mainz, D-55099, Mainz, Germany

[3] Department of Biology, Oberlin College, Science Center K111, 119 Woodland Street, Oberlin, OH 44074-1097, USA

[4] Department of Ecology & Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048, USA

[5] Department of Plant Sciences, University of Cambridge, Tennis Court Road, Cambridge, CB2 3EA, United Kingdom

[6] Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV, 89577, USA

[*] Correspondence to be sent to: Diego F. Morales-Briones and Ya Yang. Department of Plant and Microbial Biology, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108, USA, Telephone: +1 612-625-6292 (YY) Email: dfmoralesb@gmail.com; yangya@umn.edu

Abstract. **—** Gene tree discordance in large genomic datasets can be caused by evolutionary processes such as incomplete lineage sorting and hybridization, as well as model violation, and errors in data processing, orthology inference, and gene tree estimation. Species tree methods that identify and accommodate all sources of conflict are not available, but a combination of multiple approaches can help tease apart alternative sources of conflict. Here, using a phylotranscriptomic analysis in combination with reference genomes, we test a hypothesis of ancient hybridization events within the plant family Amaranthaceae s.l. that was previously supported by morphological, ecological, and Sanger-based molecular data. The dataset included seven genomes and 88 transcriptomes, 17 generated for this study. We examined gene-tree discordance using coalescent-based species trees and network inference, gene tree discordance analyses, site pattern tests of introgression, topology tests, synteny analyses, and simulations. We found that a combination of processes might have generated the high levels of gene tree discordance in the backbone of Amaranthaceae s.l. Furthermore, we found evidence that three consecutive short internal branches produce anomalous trees contributing to the discordance. Overall, our results suggest that Amaranthaceae s.l. might be a product of an ancient and rapid lineage diversification, and remains, and probably will remain, unresolved. This work highlights the potential problems of identifiability associated with the sources of gene tree discordance including, in particular, phylogenetic network methods. Our results also demonstrate the importance of thoroughly testing for multiple sources of conflict in phylogenomic analyses, especially in the context of ancient, rapid radiations. We provide several recommendations for exploring conflicting signals in such situations.

Keywords: Amaranthaceae; gene tree discordance; hybridization; incomplete lineage sorting; phylogenomics; transcriptomics; species tree; species network.

The exploration of gene tree discordance has become common in the phylogenetic era (Salichos et al. 2014; Smith et al. 2015; Huang et al. 2016; Pease et al. 2018) and is essential for understanding the underlying processes that shape the Tree of Life. Discordance among gene trees can be the product of multiple sources. These include errors and noise in data assembly and filtering, hidden paralogy, incomplete lineage sorting (ILS), gene duplication/loss (Pamilo and Nei 1988; Doyle 1992; Maddison 1997; Galtier and Daubin 2008), random noise from uninformative genes, as well as misspecified model parameters of molecular evolution such as substitutional saturation, codon usage bias, or compositional heterogeneity (Foster 2004; Cooper 2014; Cox et al. 2014; Liu et al. 2014). Among these potential sources of gene tree discordance, ILS is the most studied in the systematics literature (Edwards 2009), and several phylogenetic inference methods have been developed to accommodate ILS as the source of discordance (reviewed in Edwards et al. 2016; Mirarab et al. 2016; Xu and Yang 2016). More recently, methods that account for additional processes such as hybridization or introgression have gained attention. These include methods that estimate phylogenetic networks while accounting for ILS and hybridization simultaneously (e.g., Solís-Lemus and Ané 2016; Wen et al. 2018), and methods that detect introgression based on site patterns or phylogenetic invariants (e.g., Green et al. 2010; Durand et al. 2011; Kubatko and Chifman 2019). Frequently, multiple processes can contribute to gene tree heterogeneity (Holder et al. 2001; Buckley et al. 2006; Meyer et al. 2017; Knowles et al. 2018). However, at present, no method can estimate species trees from phylogenomic data while modeling multiple sources of conflict and heterogeneity in molecular substitution simultaneously. To overcome these limitations, the use of multiple phylogenetic tools and data partitioning schemes in phylogenomic datasets is essential to disentangle sources of gene tree heterogeneity and resolve recalcitrant relationships at deep and shallow nodes of the

Tree of Life (e.g., Alda et al. 2019; Widhelm et al. 2019; Prasanna et al. 2020; Roycroft et al. 2020).

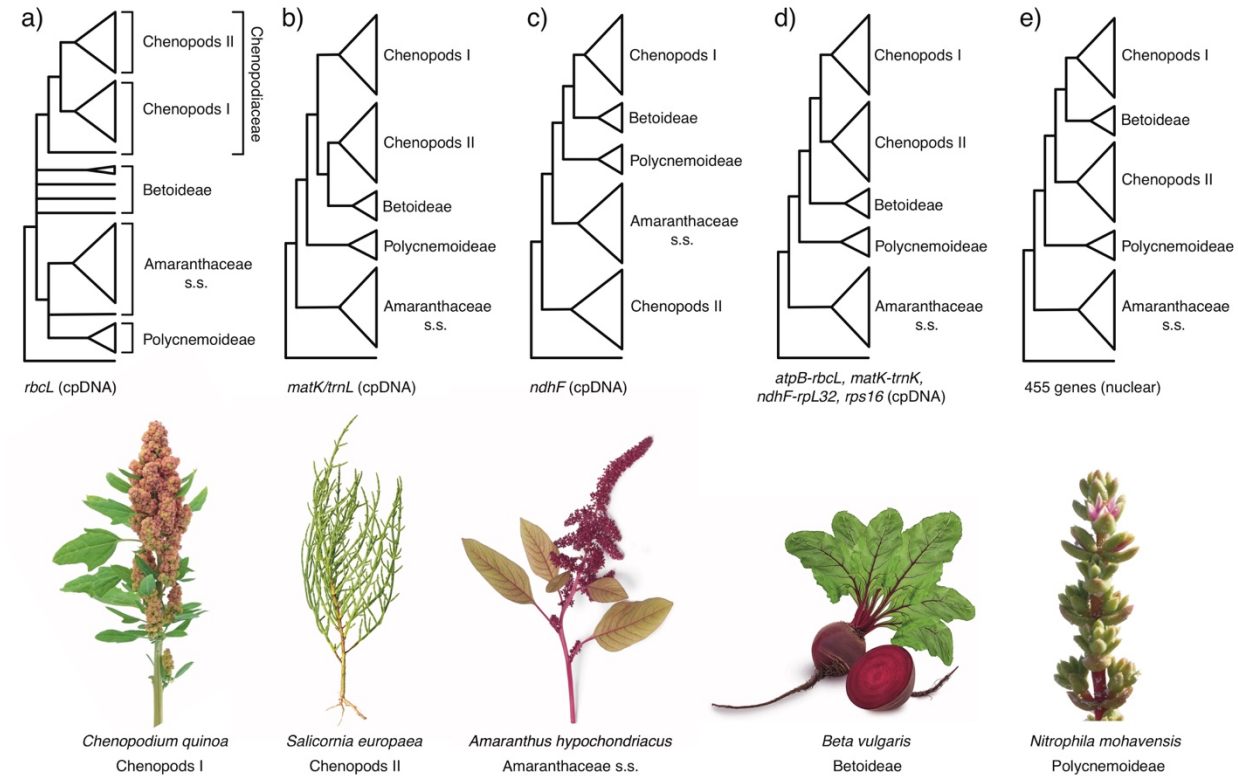
In this study, we evaluate multiple sources of gene tree conflict to test controversial hypotheses of ancient hybridization among subfamilies in the plant family Amaranthaceae s.l. Amaranthaceae s.l. includes the previously segregated family Chenopodiaceae (Hernández-Ledesma et al. 2015; The Angiosperm Phylogeny Group 2016). With ca. 2050 to 2500 species in 181 genera and a worldwide distribution (Hernández-Ledesma et al. 2015), Amaranthaceae s.l. is iconic for the repeated evolution of complex traits representing adaptations to extreme environments such as C_4 photosynthesis in hot and often dry environments (e.g., Kadereit et al. 2012; Bena et al. 2017), various modes of extreme salt tolerance (e.g., Flowers and Colmer 2015; Piirainen et al. 2017) that in several species are coupled with heavy metal tolerance (Moray et al. 2016), and very fast seed germination and production of multiple diaspore types on one individual (Kadereit et al. 2017). Several important crops are members of Amaranthaceae s.l., such as the pseudocereals quinoa and amaranth, sugar beet, spinach, glassworts, and saltworts. Many species of the family are also important fodder plants in arid regions and several are currently being investigated for their soil remediating and desalinating effects (e.g., Li et al. 2019). Due to their economic importance, reference genomes are available for *Beta vulgaris* (sugar beet, subfamily Betoideae; Dohm et al. 2014), *Chenopodium quinoa* (quinoa, Chenopodioideae; Jarvis et al. 2017), *Spinacia oleracea* (spinach; Chenopodioideae; Xu et al. 2017) and *Amaranthus hypochondriacus* (amaranth; Amaranthoideae; Lightfoot et al. 2017), representing three of the 13 currently recognized subfamilies of Amaranthaceae s.l. (sensu Kadereit et al. 2003; Kadereit et al. 2017).

Within the core Caryophyllales the previously segregated families Amaranthaceae s.s. and Chenopodiaceae have always been regarded as closely related, and their separate family status has long been the subject of phylogenetic and taxonomic debate (Kadereit et al. 2003; Masson and Kadereit 2013; Hernández-Ledesma et al. 2015; Walker et al. 2018; Fig. 1). Their close affinity is supported by a number of shared morphological, anatomical and phytochemical synapomorphies, and has been substantiated by molecular phylogenetic studies (discussed in Kadereit et al. 2003). Amaranthaceae s.s. has a predominantly tropical and subtropical distribution with the highest diversity found in the Neotropics, eastern and southern Africa and Australia (Müller and Borsch 2005), while Chenopodiaceae predominantly occurs in temperate regions and semi-arid or arid environments of subtropical regions (Kadereit et al. 2003). The key problem has always been the species-poor and heterogeneous subfamilies Polycnemoideae and Betoideae, neither of which fit easily within Chenopodiaceae or Amaranthaceae s.s. (cf. Table 5 in Kadereit et al. 2003). Polycnemoideae is similar in ecology and distribution to Chenopodiaceae but shares important floral traits such as petaloid tepals, filament tubes and 2-locular anthers with Amaranthaceae s.s. Morphologically, Betoideae fits into either Chenopodiaceae or Amaranthaceae s.s. but has a unique fruit type—a capsule that opens with a circumscissile lid (Kadereit et al. 2006). Both Betoideae and Polycnemoideae possess only a few species each and each has a strongly disjunct distribution pattern across three continents. Furthermore, the genera of both subfamilies display a number of morphologically dissociating features. Both intercontinental disjunctions of species-poor genera and unique or intermediate morphological traits led to the hypothesis that Betoideae and Polycnemoideae might have originated from hybridization events among early-branching lineages in Amaranthaceae s.l. (Hohmann et al. 2006; Masson and Kadereit 2013). To test this hypothesis, a

phylotranscriptomic approach is particularly compelling as it not only provides thousands of low-copy nuclear genes for dissecting sources of phylogenetic discordance, but also enables future studies associating gene tree topology with gene function and habitat adaptation.

Previous molecular phylogenetic analyses struggled to resolve the relationships among Betoideae, Polycnemoideae and the rest of the Amaranthaceae s.l. (Fig. 1). The first phylogenomic study of Amaranthaceae s.l. using nuclear loci (Walker et al. 2018; Fig. 1e) revealed that gene tree discordance mainly occurred at deep nodes of the phylogeny involving Betoideae. Polycnemoideae was sister to Chenopodiaceae, albeit supported by only 17% of gene trees, which contradicted previous analyses based on plastid data (Fig. 1, a–d). However, only a single species of Betoideae (the cultivated beet and its wild relative) was sampled in Walker et al. (2018). Furthermore, Walker et al. (2018) found conflicting topologies between concatenated and coalescent-based analyses, but sources of conflicting signals among gene trees remained unexplored.

In this study, we used a large genomic dataset to examine sources of gene tree discordance in Amaranthaceae s.l. Specifically, we tested whether Polycnemoideae and Betoideae result from independent hybridizations between Amaranthaceae s.s. and Chenopodioideae by distinguishing the signal of hybridization from gene tree discordance produced by ILS, uninformative gene trees, hidden paralogy, misspecifications of model of molecular evolution, and hard polytomy.



138

139 **FIGURE 1.** Phylogenetic hypothesis of Amaranthaceae s.l. from previous studies. a) Kadereit et
140 al. (2003) using the plastid (cpDNA) *rbcL* coding region. b) Müller and Borsch (2005); using the
141 cpDNA *matK* coding region and partial *trnL* intron. c) Hohmann et al. (2006) using the cpDNA
142 *ndhF* coding region. d) Kadereit et al. (2017) using the cpDNA *atpB-rbcL* spacer, *matK* with
143 *trnL* intron, *ndhF-rpL32* spacer, and *rps16* intron e) Walker et al. (2018) using 455 nuclear genes
144 from transcriptome data. Major clades of Amaranthaceae s.l. named following the results of this
145 study. Image credits: *Amaranthus hypochondriacus* by Picture Partners, *Beta vulgaris* by Olha
146 Huchek, *Chenopodium quinoa* by Diana Mower, *Nitrophila mohavensis* by James M. André, and
147 *Salsola soda* by Homeydesign.

148

149 ## MATERIALS AND METHODS

150 An overview of all dataset and phylogenetic analyses can be found in Figure S1.

151

Taxon sampling, transcriptome sequencing

We sampled 92 ingroup species (88 transcriptomes and four genomes) representing 53 genera (out of ca. 181) of all 13 currently recognized subfamilies and 16 out of 17 tribes of Amaranthaceae s.l. (sensu [Kadereit et al. 2003; Kadereit et al. 2017]). In addition, 13 outgroups across the Caryophyllales were included (ten transcriptomes and three genomes; Table S1). We generated 17 new transcriptomes for this study on an Illumina HiSeq2500 platform (Table S2). Library preparation was carried out using either poly-A enrichment or ribosomal RNA depletion. See Supplemental Methods for details on tissue collection, RNA isolation, library preparation, and quality control.

Transcriptome data processing, assembly, homology and orthology inference

Read processing, assembly, translation, and homology and orthology inference followed the 'phylogenomic dataset construction' pipeline (Yang and Smith 2014) with multiple updates. We briefly describe our procedure below, with details in the Supplemental Methods and updated scripts in https://bitbucket.org/yanglab/phylogenomic_dataset_construction/

We processed raw reads for all 88 transcriptome datasets (except *Bienertia sinuspersici*) used in this study (Table S1). Reads were corrected for errors, trimmed for sequencing adapters and low-quality bases, and filtered for organellar reads. *De novo* assembly of processed nuclear reads was carried out with Trinity v 2.5.1 (Grabherr et al. 2011) with default settings, but without *in silico* normalization. Low-quality and chimeric transcripts were removed. Filtered transcripts were clustered into putative genes with Corset v 1.07 (Davidson and Oshlack 2014) and only the longest transcript of each putative gene was retained (Chen et al. 2019). Lastly, transcripts were translated, and identical coding sequences (CDS) were removed. Homology inference was

carried out on CDS using reciprocal BLASTN, followed by orthology inference using the 'monophyletic outgroup' approach (Yang and Smith 2014), keeping only ortholog groups with at least 25 ingroup taxa.

Assessment of recombination

Coalescent species tree methods assume that there is free recombination between loci and no recombination within loci. To determine the presence of recombination in our dataset, we used the pairwise homoplasy index test Φ for recombination, as implemented in PhiPack (Bruen et al. 2006). We tested recombination on the final set of ortholog alignments (with a minimum of 25 taxa) with the default sliding window size of 100 bp. Alignments that showed a strong signal of recombination with $p \leq 0.05$ were removed from all subsequent phylogenetic analyses.

Nuclear phylogenetic analysis

We used both concatenation and coalescent-based methods to reconstruct the phylogeny of Amaranthaceae s.l. Sequences from final orthologs were aligned using MAFFT v 7.307 (Katoh and Standley 2013) with settings '—genafpair --maxiterate 1000'. Columns with more than 70% missing data were trimmed with Phyx (Brown et al. 2017), and alignments with at least 1,000 characters and 99 out of 105 taxa were retained. We first estimated a maximum likelihood (ML) tree of the concatenated matrix with RAxML v 8.2.11 (Stamatakis 2014) using a partition-by-gene scheme with GTRCAT model for each partition and clade support assessed with 200 rapid bootstrap (BS) replicates. To estimate a coalescent-based species tree, first we inferred individual ML gene trees using RAxML with a GTRCAT model and 200 BS replicates to assess clade support. Gene trees were then used to infer a species tree with ASTRAL-III v5.6.3 (Zhang et al.

198 2018) using local posterior probabilities (LPP; Sayyari and Mirarab 2016) to assess clade
199 support.
200
201 *Detecting and visualizing nuclear gene tree discordance*
202 To explore discordance among gene trees, we first calculated the internode certainty all (ICA)
203 value to quantify the degree of conflict on each node of a target tree (i.e., species tree) given
204 individual gene trees (Salichos et al. 2014). In addition, we calculated the number of conflicting
205 and concordant bipartitions on each node of the species trees. Both the ICA scores and
206 conflicting/concordant bipartitions were calculated with Phyparts (Smith et al. 2015), mapping
207 against the inferred ASTRAL species trees, using individual gene trees with BS support of at
208 least 50% for the corresponding node. Additionally, in order to distinguish strong conflict from
209 weakly supported branches, we carried out Quartet Sampling (QS; Pease et al. 2018) with 100
210 replicates. Quartet Sampling subsamples quartets from the input tree and alignment and assesses
211 the confidence, consistency, and informativeness of each internal branch by the relative
212 frequency of the three possible quartet topologies (Pease et al. 2018). Both ICA and Quartet
213 Sampling scores provide an alternative branch support that reflects underlying gene tree conflict
214 and that is not affected by anomalous high levels of bootstrap support common in phylogenomic
215 data (Kumar et al. 2012).
216 To further visualize conflict, we built a cloudogram using DensiTree v2.2.6 (Bouckaert
217 and Heled 2014). As DensiTree cannot accommodate missing taxa among gene trees, we
218 reduced the final ortholog alignments to include 41 species (38 ingroup and 3 outgroups) in order
219 to include as many orthologs as possible while representing all main clades of Amaranthaceae
220 s.l. (see results). Individual gene trees were inferred as previously described. Trees were time-

calibrated with TreePL v1.0 (Smith and O'Meara 2012) by fixing the crown age of Amaranthaceae s.l. to 66–72.1 based on a pollen record of *Polyporina cribraria* from the late Cretaceous (Maastrichtian; Srivastava 1969), and the root for the reduced 41-species dataset (most common recent ancestor of Achatocarpaceae and Aizoaceae) was set to 95 Ma based on the time-calibrated plastid phylogeny of Caryophyllales from Yao et al. (2019).

Plastid assembly and phylogenetic analysis

Although DNase treatment was carried out to remove genomic DNA, due to their high copy number, plastid sequences are often carried over in RNA-seq libraries. In addition, as young leaf tissue was used for RNA-seq, the presence of RNA from plastid genes is expected to be represented. To investigate phylogenetic signal from plastid sequences, *de novo* assemblies were carried out with the Fast-Plast v.1.2.6 pipeline (https://github.com/mrmckain/Fast-Plast) using the filtered organelle reads. Contigs produced by Spades v 3.9.0 (Bankevich et al. 2012) were mapped to the closest available reference plastomes (Table S3), one copy of the Inverted Repeat was removed, and the remaining contigs manually edited in Geneious v.11.1.5 (Kearse et al. 2012) to produce the final oriented contigs.

Contigs were aligned with MAFFT with the setting '--auto'. Two samples (*Dysphania schraderiana* and *Spinacia turkestanica*) were removed due to low sequence occupancy. Using the annotations of the reference genomes (Table S3), the coding regions of 78 genes were extracted and each gene alignment was visually inspected in Geneious to check for potential misassemblies. From each gene alignment, taxa with short sequences (i.e., < 50% of the aligned length) were removed and the remaining sequences realigned with MAFFT. The genes *rpl32* and *ycf2* were excluded from downstream analyses due to low taxon occupancy (Table S4). For each

individual gene we performed extended model selection (Kalyaanamoorthy et al. 2017) followed by ML gene tree inference and 1,000 ultrafast bootstrap replicates for branch support (Hoang and Chernomor 2018) in IQ-TREE v.1.6.1 (Nguyen et al. 2015). For the concatenated matrix we searched for the best partition scheme (Lanfear et al. 2012) followed by ML gene tree inference and 1,000 ultrafast bootstrap replicates for branch support in IQ-Tree. Additionally, we evaluated branch support with QS using 1,000 replicates and gene tree discordance with PhyParts. Lastly, to identify the origin of the plastid reads (i.e., genomic or RNA), we predicted RNA editing from CDS alignments using PREP (Mower 2009) with the alignment mode (PREP-aln), and a cutoff value of 0.8.

Species network analysis using a reduced 11-taxon dataset

We inferred species networks that model ILS and gene flow using a maximum pseudo-likelihood approach (Yu and Nakhleh 2015). Species network searches were carried out with PhyloNet v.3.6.9 (Than et al. 2008) with the command 'InferNetwork_MPL' and using the individual gene trees as input. Due to computational restrictions, and given our main focus to identify potential reticulating events among major clades of Amaranthaceae s.l., we reduced our taxon sampling to one outgroup and ten ingroup taxa to include two representative species from each of the five well-supported major lineages in Amaranthaceae s.l. (see results). We filtered the final 105-taxon ortholog set to include genes that have all 11 taxa [referred herein as 11-taxon(net) dataset; Fig S1.]. After alignment and trimming we kept genes with a minimum of 1,000 aligned base pairs and individual ML gene trees were inferred using RAxML with a GTRGAMMA model and 200 bootstrap replicates. We carried out five network searches by allowing one to five reticulation events and ten runs for each search. To estimate the optimum number of reticulations, we

optimized the branch lengths and inheritance probabilities and computed the likelihood of the best scored network from each of the five maximum reticulation events searches. Network likelihoods were estimated given the individual gene trees using the command 'CalGTProb' in PhyloNet (Yu et al. 2012). Then, we performed model selection using the bias-corrected Akaike information criterion (AICc; Sugiura 1978), and the Bayesian information criterion (BIC; Schwarz 1978). The number of parameters was set to the number of branch lengths being estimated plus the number of hybridization probabilities being estimated. The number of gene trees used to estimate the likelihood was used to correct for finite sample size. To compare network models to bifurcating trees, we also estimated bifurcating concatenated ML and coalescent-based species trees and a plastid tree as previously described with the reduced 11-species taxon sampling.

Hypothesis testing and detecting introgression using four-taxon datasets

Given the signal of multiple clades potentially involved in hybridization events detected by PhyloNet (see results), we next conducted quartet analyses to explore a single event at a time. First, we further reduced the 11-taxon(net) dataset to six taxa that included one outgroup genome (*Mesembryanthemum crystallinum*) and one ingroup from each of the five major ingroup clades: *Amaranthus hypochondriacus* (genome), *Beta vulgaris* (genome), *Chenopodium quinoa* (genome), *Caroxylon vermiculatum* (transcriptome), and *Polycnemum majus* (transcriptome) to represent Amaranthaceae s.s., Betoideae, 'Chenopods I', 'Chenopods II' and Polycnemoideae, respectively. We carried out a total of ten quartet analyses using all ten four-taxon combinations that included three out of five ingroup species and one outgroup. We filtered the final set of 105-taxon orthologs for genes with all four taxa for each combination and inferred individual gene

trees as described before. For each quartet we carried out the following analyses. We first estimated a species tree with ASTRAL and explored gene tree conflict with PhyParts. We then explored individual gene tree resolution by calculating the Tree Certainty (TC) score (Salichos et al. 2014) in RAxML using the majority rule consensus tree across the 200 bootstrap replicates. Next, we explored potential correlation between TC score and alignment length, GC content and alignment gap proportion using a linear regression model in R v.3.6.1 (R Core Team 2019). Lastly, we tested for the fit of gene trees to the three possible rooted quartet topologies for each gene using the approximately unbiased (AU) tests (Shimodaira 2002). We carried out ten constraint searches for each of three topologies in RAxML with the GTRGAMMA model, then calculated site-wise log-likelihood scores for the three constraint topologies in RAxML using GTRGAMMA and carried out the AU test using Consel v.1.20 (Shimodaira and Hasegawa 2001). In order to detect possible introgression among species of each quartet, first we estimated a species network with PhyloNet using a full maximum likelihood approach (Yu et al. 2014) with 100 runs per search while optimizing the likelihood of the branch lengths and inheritance probabilities for every proposed species network. Furthermore, we also carried out the ABBA/BABA test to detect introgression (Green et al. 2010; Durand et al. 2011) in each of four-taxon species trees. We calculated the *D*-statistic and associated *z* score for the null hypothesis of no introgression ($D = 0$) following each quartet ASTRAL species tree for taxon order assignment using 100 jackknife replicates and a block size of 10,000 bp with evobiR v1.2 (Blackmon and Adams) in R.

Additionally, to detect any non-random genomic block of particular quartet topology (Fontaine et al. 2015), we mapped the physical location of genes supporting each alternative

quartet topology onto the *Beta vulgaris* reference genome using a synteny approach (See Supplemental Information for details).

Assessment of substitutional saturation, codon usage bias, compositional heterogeneity, and model of sequence evolution misspecification

Analyses were carried out in a 11-taxon dataset [referred herein as 11-taxon(tree); Fig. S1] that included the same taxa used for species network analyses, but was processed differently to account for codon structure (see Supplemental Methods for details). Saturation was evaluated by plotting the uncorrected genetic distances of the concatenated alignment against the inferred distances (see Supplemental Methods for details). To determine the effect of saturation in the phylogenetic inferences we estimated individual ML gene trees using an unpartitioned alignment, a partition by first and second codon positions, and the third codon positions, and by removing all third codon positions. All tree searches were carried out in RAxML with a GTRGAMMA model and 200 bootstrap replicates. We then estimated a coalescent-based species tree and explored gene tree discordance with PhyParts.

Codon usage bias was evaluated using a correspondence analysis of the Relative Synonymous Codon Usage (RSCU; see Supplemental Methods for details). To determine the effect of codon usage bias in the phylogenetic inferences we estimated individual gene trees using codon-degenerated alignments (see Supplemental Methods for details). Gene tree inference and discordance analyses were carried out on the same three data schemes as previously described.

Among-lineage compositional heterogeneity was evaluated on individual genes using a compositional homogeneity test (Supplemental Methods for details). To assess if compositional

heterogeneity had an effect in species tree inference and gene tree discordance, gene trees that showed the signal of compositional heterogeneity were removed from saturation and codon usage analyses and the species tree and discordance analyses were rerun.

To explore the effect of sequence evolution model misspecification, we reanalyzed the datasets from the saturation and codon usage analyses using inferred gene trees that accounted for model selection. Additionally, we also explored saturation and model misspecification in phylogenetic trees from amino acid alignments (see Supplemental Methods for details).

Polytomy test

To test if the gene tree discordance among the main clades of Amaranthaceae s.l. could be explained by polytomies instead of bifurcating nodes, we carried out the quartet-based polytomy test by Sayyari and Mirarab (2018) as implemented in ASTRAL. We performed the polytomy test using the gene trees inferred from the saturation and codon usage analyses [11-taxon(tree) dataset]. Because this test can be sensitive to gene tree error (Syyari and Mirarab 2018), we performed a second test using gene trees where branches with less than 75% of bootstrap support were collapsed.

Coalescent simulations

To investigate if gene tree discordance can be explained by ILS alone, we carried out coalescent simulations similar to Cloutier et al. (2019). An ultrametric species tree with branch lengths in mutational units (μT) was estimated by constraining an ML tree search of the 11-taxon(net) concatenated alignment to the ASTRAL species tree topology with a GTR+GAMMA model while enforcing a strict molecular clock in PAUP v4.0a (build 165; Swofford 2002). The

mutational branch lengths from the constrained tree and branch lengths in coalescent units ($\tau = T/4N_e$) from the ASTRAL species trees were used to estimate the population size parameter theta ($\Theta = \mu T/\tau$; Degnan and Rosenberg 2009) for internal branches. Terminal branches were set with a population size parameter theta of one. We used the R package Phybase v. 1.4 (Liu and Yu 2010) that uses the formula from Rannala and Yang (2003) to simulate 10,000 gene trees using the constraint tree and the estimated theta values. Then we calculated the distribution of Robinson and Foulds (1981) tree-to-tree distances between the species tree and each gene tree using the R package Phangorn v2.5.3 (Schliep 2011), and compared this with the distribution of tree-to-tree distances between the species tree and the simulated gene tree. We ran simulations using the species tree and associated gene tree distribution from the original no partition 11-taxon(net).

Test of the anomaly zone

The anomaly zone occurs where a set of short internal branches in the species tree produces gene trees that differ from the species tree more frequently than those that are concordant [a(x); as defined in equation 4 of Degnan and Rosenberg (2006)]. To explore if gene tree discordance observed in Amaranthaceae s.l. is a product of the anomaly zone, we estimated the boundaries of the anomaly zone [a(x); as defined in equation 4 of Degnan and Rosenberg (2006)] for the internal nodes of the species tree. Here, x is the branch length in coalescent units in the species tree that has a descendant internal branch. If the length of the descendant internal branch (y) is smaller than a(x), then the internode pair is in the anomaly zone and is likely to produce anomalous gene trees (AGTs). We carried out the calculation of a(x) following Linkem et al. (2016) in the same 11-taxon(tree) ASTRAL species tree used for coalescent simulations. Additionally, to establish the frequency of gene trees that were concordant with the estimated

species trees, we quantified the frequency of all 105 possible rooted gene trees with Amaranthaceae s.l. being monophyletic.

RESULTS

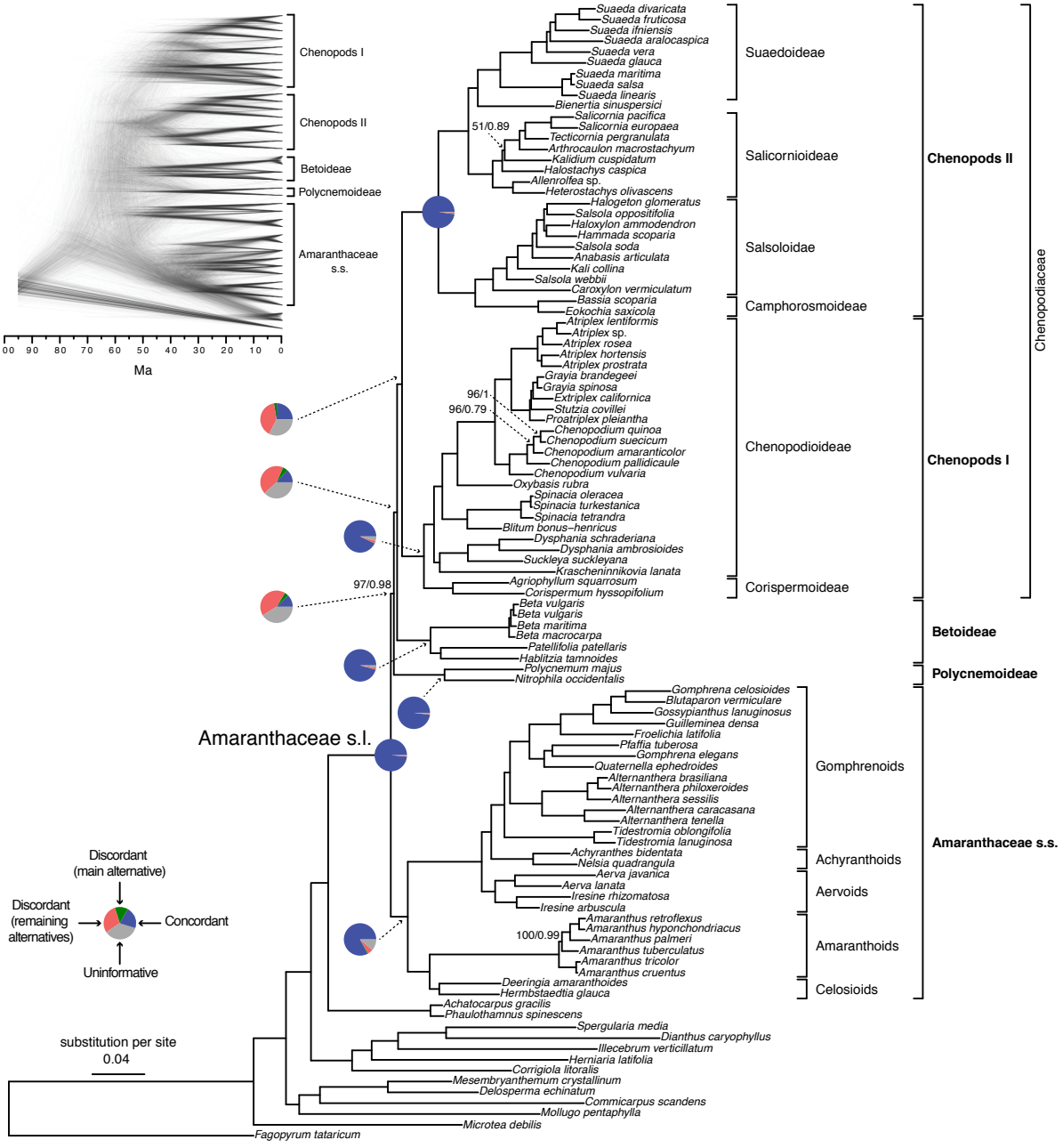
Transcriptome sequencing, assembly, translation, and quality control

Raw reads for the 17 newly generated transcriptomes are available from the NCBI Sequence Read Archive (BioProject: PRJNA640363; Table S2). The number of raw read pairs ranged from 17 to 27 million. For the 16 samples processed using RiboZero, organelle reads accounted for 15% to 52% of read pairs (Table S2). For *Tidestromia oblongifolia* that poly-A enrichment was carried out in library prep with ~5% of raw reads were from organelle (Table S2). The final number of orthologs was 13,024 with a mean of 9,813 orthologs per species (Table S1). Of those, 82 orthologs had a strong signal of recombination ($P \leq 0.05$) and were removed from downstream analyses.

Analysis of the nuclear dataset of Amaranthaceae s.l.

The final set of nuclear orthologous genes included 936 genes with at least 99 out of 105 taxa and 1,000 bp in aligned length after removal of low occupancy columns (the 105-taxon dataset). The concatenated matrix consisted of 1,712,054 columns with a gene and character occupancy of 96% and 82%, respectively. The species tree from ASTRAL and the concatenated ML tree from RAxML recovered the exact same topology with most clades having maximal support [i.e., bootstrap percentage (BS) = 100, local posterior probabilities (LPP) = 1; Fig. 2; Figs S2–S3]. Both analyses recovered Chenopodiaceae as monophyletic with the relationships among major clades concordant with the cpDNA analysis from Kadereit et al. (2017; Fig. 1d). Betoideae was

404 placed as sister of Chenopodiaceae, while Polycnemoideae was strongly supported as sister (BS
405 = 97, LPP = 0.98) to the clade composed of Chenopodiaceae and Betoideae. Amaranthaceae s.s.
406 had an overall topology concordant to Kadereit et al. (2017), with the exception of *Iresine,* which
407 was recovered among the Aervoids (Fig. 2; Figs S2–S3).
408



409

FIGURE 2. Maximum likelihood phylogeny of Amaranthaceae s.l. inferred from RAxML analysis of the concatenated 936-nuclear gene supermatrix, which had the same topology as recovered from ASTRAL. All nodes have maximal support (bootstrap = 100/ASTRAL local posterior probability = 1) unless noted. Pie charts present the proportion of gene trees that support that clade (blue), support the main alternative bifurcation (green), support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Only pie charts for major clades are shown (see Fig. S2 for all node pie charts). Branch lengths are in number of substitutions per site. The inset (top left) shows the Densitree cloudogram inferred from 1,242 nuclear genes for the reduced 41-taxon dataset.

The conflict analyses confirmed the monophyly of Amaranthaceae s.l. with 922 out of 930 informative gene trees being concordant (ICA= 0.94) and having full QS support (1/–/1; i.e., all sampled quartets supported that branch). Similarly, the monophyly of Amaranthaceae s.s. was highly supported by 755 of 809 informative gene trees (ICA =0.85) and the QS scores (0.92/0/1). However, the backbone of the family was characterized by high levels of gene tree discordance (Fig. 2; Figs S2–S3). The monophyly of Chenopodiaceae was supported only by 231 out of 632 informative gene trees (ICA = 0.42) and the QS score (0.25/0.19/0.99) suggested weak quartet support with a skewed frequency for an alternative placement of two well-defined clades within Chenopodiaceae, herein referred to as ‘Chenopods I’ and ‘Chenopods II’ (Fig. 2; Figs S2–S3). ‘Chenopods I’ and ‘Chenopods II’ were each supported by the majority of gene trees, 870 (ICA = 0.89) and 916 (ICA = 0.91), respectively and full QS support. Similarly, high levels of conflict among informative gene trees were detected in the placement of Betoideae (126 out of 579 informative genes being concordant, ICA = 0.28; QS score 0.31/0.57/1) and Polycnemoideae (116/511; ICA = 0.29;0.3/0.81/0.99). The Densitree cloudogram also showed significant conflict along the backbone of Amaranthaceae s.l. (Fig. 2).

Together, analysis of nuclear genes recovered five well-supported clades in Amaranthaceae s.l.: Amaranthaceae s.s., Betoideae, 'Chenopods I', 'Chenopods II', and Polycnemoideae. However, relationships among these five clades showed a high level of conflict among genes [ICA scores and gene counts (pie charts)] and among subsampled quartets (QS scores), despite having high support from both BS and LPP scores.

Plastid phylogenetic analysis of Amaranthaceae s.l.

RNA editing prediction analysis revealed editing sites only on CDS sequences of reference plastomes (Table S3), suggesting that cpDNA reads in RNA-seq libraries come from RNA rather than DNA leftover from incomplete DNase digestion during sample processing (See Discussion for details in plastid assembly from RNA-seq data).

The final alignment from 76 genes included 103 taxa and 55,517 bp in aligned length. The ML tree recovered the same five main clades within Amaranthaceae s.l. with maximal support (BS = 100; Figs S4–S6). Within each main clade, relationships were fully congruent with Kadereit et al. (2017) and mostly congruent with our nuclear analyses. However, the relationship among the five main clades differed from the nuclear tree. Here, the sister relationships between Betoideae and 'Chenopods I', and between Amaranthaceae s.s. and Polycnemoideae were both supported by BS =100. The sister relationship between these two larger clades was moderately supported (BS = 73), leaving 'Chenopods II' as sister to the rest of Amaranthaceae s.l.

Conflict analysis confirmed the monophyly of Amaranthaceae s.l. with 51 out of 69 informative gene trees supporting this clade (ICA = 0.29) and full QS support (1/–/1). On the other hand, and similar to the nuclear phylogeny, significant gene tree discordance was detected

among plastid genes regarding placement of the five major clades (Figs S4–S6). The sister relationship of Betoideae and 'Chenopods I' was supported by only 20 gene trees (ICA = 0.06), but it had a strong support from QS (0.84/0.88/0/94). The relationship between Amaranthaceae s.s. and Polycnemoideae was supported by only 15 gene trees (ICA = 0.07), while QS showed weak support (0.41/0.21.0.78) with signals of a supported secondary evolutionary history. The clade uniting Betoideae, 'Chenopods I', Amaranthaceae s.s., and Polycnemoideae was supported by only four-gene trees, with counter-support from both QS (-0.29/0.42/0.75) and ICA (-0.03), suggesting that most gene trees and sampled quartets supported alternative topologies.

Species network analysis of Amaranthaceae s.l.

The reduced 11-taxon(net) dataset included 4,138 orthologous gene alignments with no missing taxon and a minimum of 1,000 bp (aligned length after removal of low occupancy columns). The 11-taxon(net) ASTRAL species tree was congruent with the 105-taxon tree, while both the nuclear and plastid ML trees from concatenated supermatrices had different topologies than their corresponding 105-taxon trees (Fig. 3). Model selection indicated that any species network was a better model than the best bifurcating nuclear or plastid trees (ASTRAL; AICc = 46972.9794; Table S5). PhyloNet identified up to five hybridization events among the clades of Amaranthaceae s.l. (Fig. 3), with the best model having five hybridization events involving all five clades (AICc = 28459.1835; Table S5). The best species network did not support the hypothesis of the hybrid origin of Betoideae or Polycnemoideae. Moreover, the best species network showed a complex reticulate pattern that involved mainly ‘Chenopods I’ and ‘Chenopods II’ (Fig. 3), but none of these reticulations events were supported by *D*-Statistic or species network results from the four-taxon analyses (see below).

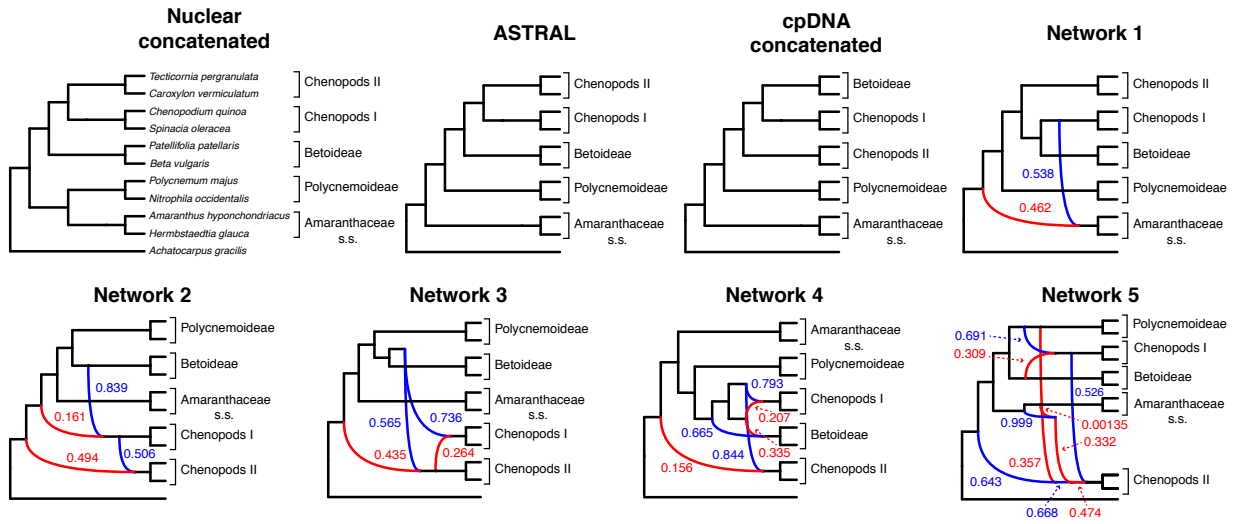


FIGURE 3. Species trees and species networks of the reduced 11-taxon(net) dataset of Amaranthaceae s.l. Nuclear concatenated phylogeny inferred from 4,138-nuclear gene supermatrix with RAxML. ASTRAL species tree inferred using 4,138 nuclear genes. cpDNA concatenated tree inferred from 76-plastid gene supermatrix with IQ-tree. Species network inferred from PhyloNet pseudolikelihood analyses with 1 to 5 maximum number of reticulations. Red and blue indicate the minor and major edges, respectively, of hybrid nodes. Number next to the branches indicates inheritance probabilities for each hybrid node.

Four-taxon analyses

To test for hybridization events one at a time, we further reduced the 11-taxon(net) dataset to 10 four-taxon combinations that each included one outgroup and one representative each from three out of the five major ingroup clades. Between 7,756 and 8,793 genes were used for each quartet analysis (Table S6) and each quartet topology can be found in Figure 4. Only five out of the ten bifurcating quartet species trees (H0 and more frequent gene tree) were compatible with the nuclear species tree inferred from the complete 105-taxon dataset. The remaining quartets corresponded to the second most frequent gene tree topology in the 105-taxon nuclear tree,

except for the quartet of Betoideae, 'Chenopods II' and Polycnemoideae (PBC2, which correspond to the least frequent gene tree).

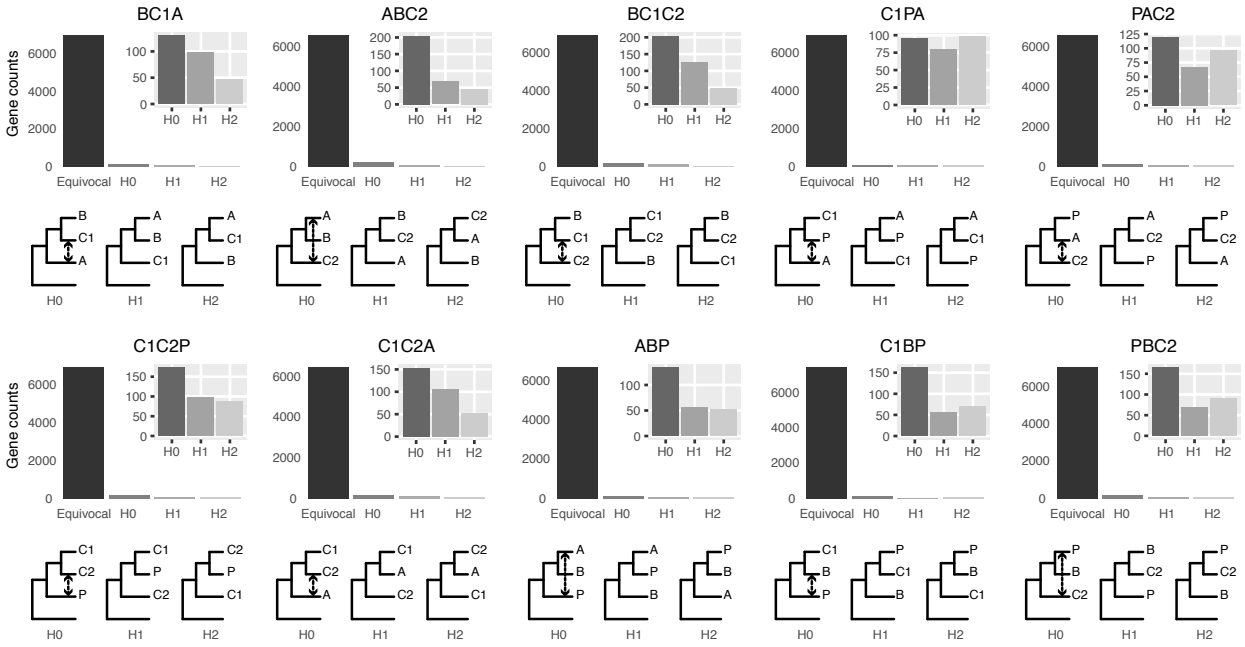


FIGURE 4. Gene counts from Approximate-Unbiased (AU) topology test of the 10 quartets from the five main clades of Amaranthaceae s.l. AU tests were carried out between the three possible topologies of each quartet. H0 represents the ASTRAL species tree of each quartet. "Equivocal" indicates gene trees that fail to reject all three alternative topologies for a quartet with $p \leq 0.05$. Gene counts for each of the three alternative topologies represent gene trees supporting unequivocally one topology by rejecting the other two alternatives with $p \leq 0.05$. Insets represent gene counts only for unequivocal topology support. Double arrowed lines in each H0 quartet represent the direction of introgression from the ABBA/BABA test. Each quartet is named following the species tree topology, where the first two species are sister to each other. A = Amaranthaceae s.s. (represented by *Amaranthus hypochondriacus*), B = Betoideae (*Beta vulgaris*), C1 = Chenopods I (*Chenopodium quinoa*), C2 = Chenopods II (*Caroxylum vermiculatum*), P = Polycnemoideae (*Polycnemum majus*). All quartets are rooted with *Mesembryanthemum crystallinum.*

In each of the ten quartets, the ASTRAL species tree topology (H0) was the most frequent among individual gene trees (raw counts) but only accounted for 35%–41% of gene trees, with the other two alternative topologies having balanced to slightly skewed frequencies (Fig. S7a; Table S7). Gene counts based on the raw likelihood scores from the constraint analyses showed similar patterns (Fig. S7b; Table S7). When filtered by significant likelihood support (i.e., $\Delta AICc \geq 2$), the number of trees supporting each of the three possible topologies dropped between 34% and 45%, but the species tree remained the most frequent topology for all quartets (Fig. S7b; Table S7). The AU topology tests failed to reject ($P \leq 0.05$) approximately 85% of the gene trees for any of the three possible quartet topologies and rejected all but a single topology in only 3%–4.5% of cases. Among the unequivocally selected gene trees, the frequencies among the three alternative topologies were similar to ones based on raw likelihood scores (Fig S7; Table S7). Therefore, topology tests showed that most genes were uninformative for resolving the relationships among the major groups of Amaranthaceae s.l.

Across all ten quartets we found that most genes had very low TC scores (for any single node the maximum TC value is 1; Supplemental Fig. S8), showing that individual gene trees also had high levels of conflict among bootstrap replicates, which also indicated uninformative genes and was concordant with the AU topology test results. We were unable to detect any significant correlation between TC scores and alignment length, GC content or alignment gap fraction (Table S8), suggesting that filtering genes by any of these criteria was unlikely to increase the information content of the dataset.

Species network analyses followed by model selection using each of the four-taxon datasets showed that in seven out of the ten total quartets, the network with one hybridization event was a better model than any bifurcating tree topology. However, each of the best three

539 networks from PhyloNet had very close likelihood scores and no significant ΔAICc among them
540 (Table S6; Fig S9). For the remaining three quartets, the species trees (H0) was the best model.
541 The ABBA/BABA test results showed a significant signal of introgression within each of
542 the ten quartets (Table S9; Fig 4). The possible introgression was detected between six out of the
543 ten possible pairs of taxa. Potential introgression between Betoideae and Amaranthaceae s.s.,
544 'Chenopods I' or 'Chenopods II', and between 'Chenopods I' and Polycnemoideae was not
545 detected.
546 To further evaluate whether alternative quartets were randomly distributed across the
547 genome, we mapped topologies from the quartet of Betoideae, 'Chenopods II, and
548 Amaranthaceae s.s. (BC1A) onto the reference genome of *Beta vulgaris.* We used the BC1A
549 quartet as an example as all four species in this quartet have reference genomes. Synteny analysis
550 between the diploid ingroup reference genome *Beta vulgaris* and the diploid outgroup reference
551 genome *Mesembryanthemum crystallinum* recovered 22,179 collinear genes in 516 syntenic
552 blocks. With the collinear ortholog pair information, we found that of the 8,258 orthologs of the
553 BC1A quartet, 6,941 contained syntenic orthologous genes within 383 syntenic blocks. The
554 distribution of the BC1A quartet topologies along the chromosomes of *Beta vulgaris* did not
555 reveal any spatial clustering of any particular topology along the chromosomes (Fig. S10).
556 Gene Ontology enrichment analyses (not shown) using alternative topologies of the
557 BC1A quartet did not recover any significant term associated with C_4 photosynthesis, drought
558 recovery, or salt stress response.
559
560

Assessment of substitutional saturation, codon usage bias, compositional heterogeneity, sequence evolution model misspecification, and polytomy test.

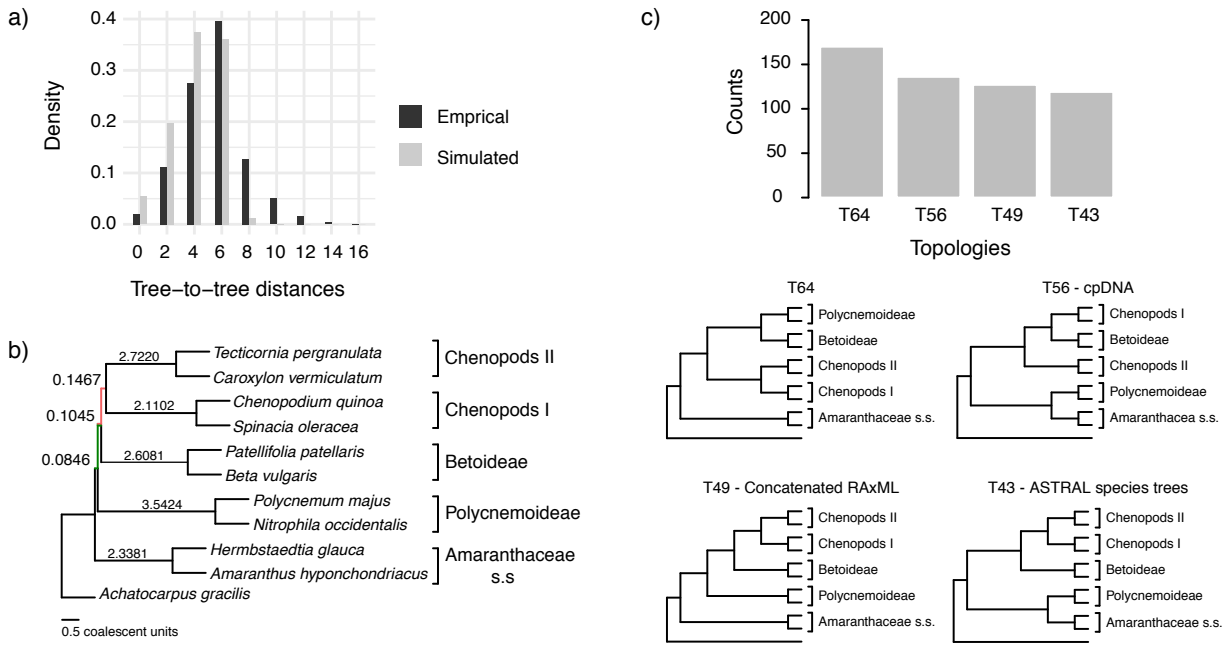
We assembled a second 11-taxon(tree) dataset that included 5,936 genes and a minimum of 300 bp (aligned length after removal of low occupancy columns) and no missing taxon. The saturation plots of uncorrected and predicted genetic distances showed that the first and second codon positions were unsaturated ($y = 0.884x$), whereas the slope of the third codon positions ($y = 0.571x$) showed a signal of saturation (Fig. S11). The correspondence analyses of RSCU show that some codons are more frequently used in different species, but overall the codon usage was randomly dispersed among all species and not clustered by clade (Fig. S12). This suggests that the phylogenetic signal is unlikely to be driven by differences in codon usage bias among clades. Furthermore, only 549 (~9%) genes showed a signal of compositional heterogeneity ($p < 0.05$). The topology and support (LPP = 1.0) for all branches was the same for the ASTRAL species trees obtained from the different data schemes while accounting for saturation, codon usage, compositional heterogeneity, and model of sequence evolution, and was also congruent with the ASTRAL species tree and concatenated ML from the 105-taxon analyses (Fig. S13). In general, the proportion of gene trees supporting each bipartition remained the same in every analysis and showed high levels of conflict among the five major clades of Amaranthaceae s.l. (Fig S13).

The ASTRAL polytomy test resulted in the same bifurcating species tree for the 11-taxon(tree) dataset and rejected the null hypothesis that any branch is a polytomy ($p < 0.01$ in all cases). These results were identical when using gene trees with collapsed branches.

582 *Coalescent simulations and tests of the anomaly zone*

583 The distribution of tree-to-tree distances of the empirical and simulated gene trees to the species
584 tree from the 11-taxon(tree) dataset largely overlapped (Fig 5a), suggesting that ILS alone is able
585 to explain most of the observed gene tree heterogeneity (Maureira-Butler et al. 2008). The
586 anomaly zone limit calculations using species trees from the 11-taxon(tree) dataset detected two
587 pairs of internodes among the five major groups in Amaranthaceae s.l. that fell into the anomaly
588 zone (the red pair and the green pair, Fig. 5b; Table S10). Furthermore, gene tree counts showed
589 that the species tree was not the most common gene tree topology, as defined for the anomaly
590 zone (Degnan and Rosenberg 2006; Fig 5c). The species tree was the fourth most common gene
591 tree topology (119 out of 4,425 gene trees), while the three most common gene tree topologies
592 occurred 170, 136, and 127 times (Fig. 5c).

593



594

595 **FIGURE 5.** Coalescent simulations and tests of the anomaly zone from the 11-taxon(tree) dataset
596 estimated from individual gene trees. a) Distribution of tree-to-tree distances between empirical

597 gene trees and the ASTRAL species tree, compared to those from the coalescent simulation. b)
598 ASTRAL species tree showing branch length in coalescent units. Green and red branches
599 represent the internodes that fall in the anomaly zone (see Table S10 for anomaly zone limits). c)
600 Gene tree counts (top) of the four most common topologies (bottom). Gene trees that do not
601 support the monophyly of any of the five major clades were ignored.
602

603 ## DISCUSSION

604 The exploration of gene tree discordance has become a fundamental step to understand
605 recalcitrant relationships across the Tree of Life. Recently, new tools have been developed to
606 identify and visualize gene tree discordance (e.g., Salichos et al. 2014; Smith et al. 2015; Huang
607 et al. 2016; Pease et al. 2018). However, downstream methods that evaluate processes generating
608 observed patterns of gene tree discordance are still in their infancy. In this study, by combining
609 transcriptomes and genomes, we were able to create a rich and dense dataset to start to tease
610 apart alternative hypotheses concerning the sources of conflict along the backbone phylogeny of
611 Amaranthaceae s.l. We found that gene tree heterogeneity observed in Amaranthaceae s.l. can be
612 explained by a combination of processes, including ILS, ancient hybridization, and
613 uninformative genes, that might have acted simultaneously and/or cumulatively.
614

615 ### *Gene tree discordance detected among plastid genes*

616 Although both our concatenation-based plastid and nuclear phylogenies supported the same five
617 major clades of Amaranthaceae s.l., the relationships among these clades are incongruent (Figs. 2
618 & S4). Cytonuclear discordance is well-known in plants and it has been traditionally attributed to
619 reticulate evolution (Rieseberg and Soltis 1991; Sang et al. 1995; Soltis and Kuzoff 1995). Such
620 discordance continues to be treated as evidence in support of hybridization in more recent
621 phylogenomic studies that assume the plastome to be a single, linked locus (e.g., Folk et al.

2017; Vargas et al. 2017; Morales-Briones et al. 2018b; Lee-Yaw et al. 2019). However, recent work showed that the plastome might not necessarily act as a single locus and high levels of tree conflict have been detected (Gonçalves et al. 2019; Walker et al. 2019).

In Amaranthaceae s.l., previous studies based on plastid protein-coding genes or introns (Fig. 1; Kadereit et al. 2003; Müller and Borsch 2005; Hohmann et al. 2006; Kadereit et al. 2017) resulted in different relationships among the five main clades and none in agreement with our 76-gene plastid phylogeny. Our conflict and QS analyses of the plastid dataset (Figs S5–S6) revealed strong signals of gene tree discordance among the five major clades of Amaranthaceae s.l., likely due to heteroplasmy, although the exact sources of conflict are yet to be clarified (Gonçalves et al. 2019). Unlike the results found by Walker et al. (2019), our individual plastid gene trees had highly supported nodes (i.e., BS $\geq$ 70, Fig S5), suggesting that low phylogenetic information content is not the main source of conflict in our plastid dataset.

Our results support previous studies showing RNA-seq data can be a reliable source for plastome assembly (Smith 2013; Osuna-Mascaró et al. 2018; Gitzendanner et al. 2018). RNA-seq libraries can contain some genomic DNA due to incomplete digestion during RNA purification (Smith 2013). Given the AT-rich nature of plastomes, plastid DNA may survive the poly-A selection during mRNA enrichment (Schliesky et al. 2012). However, RNA editing prediction results showed that our Amaranthaceae s.l. cpDNA assemblies came from RNA rather than DNA contamination regardless of library preparation by poly-A enrichment (71 transcriptomes) or RiboZero (16 transcriptomes). Similarly, Osuna-Mascaró et al. (2018) also found highly similar plastome assemblies (i.e., general genome structure, and gene number and composition) from RNA-seq and genomic libraries, supporting the idea that plastomes are fully transcribed in photosynthetic eukaryotes (Shi et al. 2016). Furthermore, the backbone topology

of our plastid tree built mainly from RNA-seq data (97 out of 105 samples) was consistent with a recent complete plastome phylogeny of Caryophyllales mainly from genomic DNA (Yao et al. 2019), showing the utility of recovering plastid gene sequences from RNA-seq data. Nonetheless, RNA editing might be problematic when combining samples from RNA-seq and genomic DNA, especially when resolving phylogenetic relationships among closely related species.

Identifiability in methods for detecting reticulation events

All methods that we used to detect ancient hybridization inferred the presence of reticulation events. However, our results suggest that these methods all struggle with ancient, rapid radiations. Advances have been made in recent years in developing methods to infer species networks in the presence of ILS (reviewed in Elworth et al. 2019). These methods have been increasingly used in phylogenetic studies (e.g., Wen et al. 2016; Copetti et al. 2017; Morales-Briones et al. 2018a; Crowl et al. 2020). To date, however, species network inference is still computationally intensive and limited to a small number of species and a few hybridization events (Hejase and Liu 2016; but see Hejase et al. 2018 and Zhu et al. 2019). Furthermore, studies evaluating the performance of different phylogenetic network inference approaches are scarce and restricted to simple hybridization scenarios. Kamneva and Rosenberg (2017) showed that likelihood methods like Yu et al. (2014) are often robust to ILS and gene tree error when symmetric hybridization (equal genetic contribution of both parents) events are considered. While this approach usually does not overestimate hybridization events, it fails to detect skewed hybridization (unequal genetic contribution of both parents) events in the presence of significant ILS. Methods developed to scale to larger numbers of species and hybridizations like the ones

using pseudo-likelihood approximations (i.e., Solís-Lemus and Ané 2016; Yu and Nakhleh 2015) are yet to be evaluated independently, but in the case of the Yu and Nakhleh (2015) method based on rooted triples, it cannot distinguish the correct network when other networks can produce the same set of triples (Yu and Nakhleh 2015). On the other hand, the method of Solís-Lemus and Ané (2016), based on unrooted quartets, is better at avoiding indistinguishable networks, but it is limited to only level-1 network scenarios.

Applying the above methods to our data set recovered multiple reticulation events. Analysis of our 11-taxon(net) dataset using a pseudo-likelihood approach detected up to five hybridization events involving all five major clades of Amaranthaceae s.l. (Fig. 3). Model selection, after calculating the full likelihood of the obtained networks, also chose the 5-reticulation species as the best model. Likewise, we found that any species network had a better ML score than a bifurcating tree (Table S5). However, further analyses demonstrated that full likelihood network searches with up to one hybridization event are indistinguishable from each other (Table S6), resembling a random gene tree distribution. This pattern can probably be explained by the high levels of gene tree discordance and lack of phylogenetic signal in the inferred quartet gene trees (Fig. 4), suggesting that the 11-taxon(net) network searches can potentially overestimate reticulation events due to high levels of gene tree error or ILS.

Using the *D*-Statistic (Green et al. 2010; Durand et al. 2011) we also detected signals of introgression in seven possible locations among the five main groups of Amaranthaceae s.l. (Table S9). The inferred introgression events agreed with at least one of the reticulation scenarios from the phylogenetic network analysis. However, the *D*-Statistic did not detect any introgression that involves Betoideae, which was detected in the phylogenetic network analysis with either four or five reticulations events. The *D*-Statistic has been shown to be robust to a

691 wide range of divergence times, but it is sensitive to relative population size (Zheng and Janke
692 2018), which agrees with the notion that large effective population sizes and short branches
693 increase the chances of ILS (Pamilo and Nei 1988) and in turn can dilute the signal for the *D*-
694 Statistic (Zheng and Janke 2018). Recently, Elworth et al. (2018) found that multiple or 'hidden'
695 reticulations can cause the signal of the *D*-statistic to be lost or distorted. Furthermore, when
696 multiple reticulations are present, the traditional approach of dividing datasets into quartets can
697 be problematic as it largely underestimates *D* values (Elworth et al. 2018). Given short internal
698 branches in the backbone of Amaranthaceae s.l. and the phylogenetic network results showing
699 multiple hybridizations, it is plausible that our *D*-statistic may be affected by these issues.
700 Our analysis highlights problems with identifiability in relying on *D*-statistic or
701 phylogenetic network analysis alone to detect reticulation events, especially in cases of ancient
702 and rapid diversification. Both analyses resulted in highly complex and inconsistent reticulate
703 scenarios that cannot be distinguished from ILS or gene tree error. Hence, despite the use of
704 genome-scale data and exhaustive hypothesis testing, support is lacking for the hybrid origin of
705 Polycnemoideae or Betoideae, or any particular hybridization event among major groups in
706 Amaranthaceae s.l. In addition to potential hybridization events, rapid speciation, short branches,
707 and large ancestral population size all impacting our ability to resolve relationships among major
708 clades in Amaranthaceae s.l. Simulating combinations of these scenarios is beyond the scope of
709 this particular manuscript.
710
711 *ILS and the Anomaly Zone*
712 ILS is ubiquitous in multi-locus phylogenetic datasets. In its most severe cases ILS produces the
713 'anomaly zone', defined as a set of short internal branches in the species tree that produce

anomalous gene trees (AGTs) that are more likely than the gene tree that matches the species tree (Degnan and Rosenberg 2006). Rosenberg (2013) expanded the definition of the anomaly zone to require that a species tree contain two consecutive internal branches in an ancestor–descendant relationship in order to produce AGTs. To date, only a few empirical examples of the anomaly zone have been reported (Linkem et al. 2016; Cloutier et al. 2019). Our results show that the species tree of Amaranthaceae s.l. has three consecutive short internal branches that lay within the limits of the anomaly zone (i.e., $y < a[x]$; Fig. 5; Table S10) and that the species tree is not the most frequent gene tree (Fig. 4). While both lines of evidence support the presence of AGTs, it is important to point out that our quartet analysis showed that most quartet gene trees were equivocal (94–96%; Fig. 4), and therefore, were uninformative. Huang and Knowles (2009) pointed out that the gene tree discordance produced from the anomaly zone can be produced by uninformative gene trees and that for species trees with short branches the most probable gene tree topology is a polytomy rather than an AGT. Our ASTRAL polytomy test, however, rejected a polytomy along the backbone of Amaranthaceae s.l. in any of the gene tree sets used. While we did not test for polytomies in individual gene trees, our ASTRAL polytomy test using gene trees with branches of <75% bootstrap support collapsed also rejected the presence of a polytomy. Therefore, the distribution of gene tree frequency in combination with short internal branches in the species tree supports the presence of an anomaly zone in Amaranthaceae s.l.

Taxonomic implications

Despite the strong signal of gene tree discordance, both nuclear and plastid datasets strongly supported five major clades within Amaranthaceae s.l.: Amaranthaceae s.s, ‘Chenopods I’, ‘Chenopods II’, Betoideae, and Polycnemoideae (Figs. 2 & S4). These five clades are congruent

with morphology and previous taxonomic treatments of the group. However, the relationships among these five lineages remain elusive with our data. Taken together, our tests of sources of incongruence for these early-diverging nodes indicate that no single source such as a particular ancient hybridization event can confidently account for the strong signal of gene tree discordance, suggesting that the discordance results primarily from ancient and rapid lineage diversification. Thus, the backbone of Amaranthaceae s.l. remains, and likely will remain, unresolved even with genome-scale data. The stem age of Amaranthaceae s.l. dates back to the early Tertiary (Paleocene; Kadereit et al. 2012; Di Vincenzo et al. 2018; Yao et al. 2019), but due to nuclear and plastid gene tree along the backbone, the geographic origin of Amaranthaceae s.l. remains ambiguous.

Therefore, for the sake of taxonomic stability, we suggest retaining Amaranthaceae s.l. sensu APG IV (The Angiosperm Phylogeny Group 2016), which includes the previously recognized Chenopodiaceae. Amaranthaceae s.l. is characterized by a long list of anatomical, morphological and phytochemical characters such as minute sessile flowers with five tepals, a single whorl of epitepalous stamens, and one basal ovule (Kadereit et al. 2003). Here, we recognize five subfamilies within Amaranthaceae s.l. represented by the five well-supported major clades recovered in this study (Fig. 2): Amaranthoideae (Amaranthaceae s.s.), Betoideae, Chenopodioideae (‘Chenopods I’), Polycnemoideae, and Salicornioideae (‘Chenopods II’).

Conclusions

Our analyses highlight the need to test for multiple sources of conflict in phylogenomic analyses, especially when trying to resolve phylogenetic relationships with extensive phylogenetic conflict. Furthermore, one needs to be aware of the strengths and limitations of

different phylogenetic methods and be cautious about relying on any single analysis, for example in the usage of phylogenetics species networks over coalescent-based species trees (Blair and Ané 2020). We make the following recommendation on five essential steps towards exploring heterogeneous phylogenetic signals in phylogenomic datasets in general. 1) Study design: consider whether the taxon sampling and marker choice enable testing alternative sources of conflicting phylogenetic signal. For example, will there be sufficient phylogenetic signal and sufficient taxon coverage in individual gene trees for methods such as phylogenetic network analyses? 2) Data processing: care should be taken in data cleaning, partitioning (e.g., nuclear vs. plastid), and using orthology inference methods that explicitly address paralogy issues (e.g., tree-based orthology inference and synteny information). 3) Species tree inference: select species tree methods that accommodate the dataset size and data type (e.g., ASTRAL for gene tree-based inferences or SVDquartet [Chifman and Kubatko 2014] for SNP-based inferences), followed by visualization of phylogenetic conflict using tools such as the pie charts (e.g., PhyParts) and quartet-based tools (e.g., Quartet Sampling; Quadripartition Internode Certainty [Zhou et al. 2020]; Concordance Factors [Minh et al. 2020]). 4) Assessing hybridization: if phylogenetic conflict cannot be explained by processes like ILS, phylogenetic species network analyses (e.g., PhyloNet) reduced taxon sampling can be applied to test hybridization hypotheses given results in step 3; 5) Hypothesis testing: additional tests can be performed given the results of recommendation 3 and 4 and depending on the scenario. These could include testing for model misspecification, anomaly zone, uninformative gene tree, and if hybridization is hypothesized, testing putative reticulation events one at a time, as illustrated in this study.

Despite using genome-scale data and exhaustive hypothesis testing, the backbone phylogeny of Amaranthaceae s.l. remains unresolved, and we were unable to distinguish ancient

783 hybridization events from ILS or uninformative gene trees. Similar situations might not be
784 atypical across the Tree of Life. As we leverage more genomic data and explore gene tree
785 discordance in more detail, these steps will be informative in other clades, especially in those
786 that are products of ancient and rapid lineage diversification (e.g., Widhelm et al. 2019; Koenen
787 et al. 2020). Ultimately, such endeavors will be instrumental in gaining a full understanding of
788 the complexity of the Tree of Life.
789

790 ## SUPPLEMENTARY MATERIAL

791 Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/.[NNNN]
792

793 ## ACKNOWLEDGMENTS

794 The authors thank H. Freitag, J.M. Bena and the Millennium Seed Bank for providing seeds; U.
795 Martiné for assisting with RNA extraction; N. Wang and Y.-Y. Huang for sample sequencing. A.
796 Crum, R. Ree, B. Carstens, and three anonymous reviewers for providing helpful comments; the
797 Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing access
798 to computational resources. This work was supported by the University of Minnesota, the
799 University of Michigan, the US National Science Foundation (DEB 1354048), and the
800 Department of Energy, Office of Science, Genomic Science Program (Contract Number DE-
801 SC0008834).
802

REFERENCES

Alda F., Tagliacollo V.A., Bernt M.J., Waltz B.T., Ludt W.B., Faircloth B.C., Alfaro M.E., Albert J.S., Chakrabarty P. 2019. Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. Syst. Biol. 68:573–593.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.

Bena M.J., Acosta J.M., Aagesen L. 2017. Macroclimatic niche limits and the evolution of C_4 photosynthesis in Gomphrenoideae (Amaranthaceae). Bot. J. Linn. Soc. 184:283–297.

Blackmon H., Adams R.A. 2015 EvobiR: Tools for comparative analyses and teaching evolutionary biology. doi:10.5281/zenodo.30938

Blair C., Ané C. 2020. Phylogenetic Trees and Networks Can Serve as Powerful and Complementary Approaches for Analysis of Genomic Data. Syst. Biol. 69:593–601.

Bouckaert R., Heled J. 2014. DensiTree 2: Seeing Trees Through the Forest. BioRxiv. 012401.

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx - phylogenetic tools for unix. Bioinformatics. 33:1886–1888.

Bruen T.C., Philippe H., Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics. 172:2665–2681.

Buckley T.R., Cordeiro M., Marshall D.C., Simon C. 2006. Differentiating between Hypotheses of Lineage Sorting and Introgression in New Zealand Alpine Cicadas (Maoricicada Dugdale). Syst. Biol. 55:411–425.

826 Chen L.-Y., Morales-Briones D.F., Passow C.N., Yang Y. 2019. Performance of gene expression
827 analyses using de novo assembled transcripts in polyploid species. Bioinformatics.
828 35:4314–4320.
829 Chifman J., Kubatko L. 2014. Quartet Inference from SNP Data Under the Coalescent Model.
830 Bioinformatics. 30:3317–3324.
831 Cloutier A., Sackton T.B., Grayson P., Clamp M., Baker A.J., Edwards S.V. 2019. Whole-
832 Genome Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the
833 Presence of an Empirical Anomaly Zone. Syst. Biol. 68:937–955
834 Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends
835 Plant Sci. 19:576–582.
836 Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S.,
837 Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson
838 M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North
839 American columnar cacti. Proc. Natl. Acad. Sci. 114:12003–12008.
840 Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting Phylogenies for Early
841 Land Plants are Caused by Composition Biases among Synonymous Substitutions. Syst.
842 Biol. 63:272–279.
843 Crowl A.A., Manos P.S., McVay J.D., Lemmon A.R., Lemmon E.M., Hipp A.L. 2020.
844 Uncovering the genomic signature of ancient introgression between white oak lineages
845 (*Quercus*). New Phytol. 226:1158–1170.
846 Davidson N.M., Oshlack A. 2014. Corset: enabling differential gene expression analysis for de
847 novo assembled transcriptomes. Genome Biol. 15:57.

848 Degnan J.H., Rosenberg N.A. 2006. Discordance of Species Trees with Their Most Likely Gene
849 Trees. PLoS Genet. 2:e68.
850 Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the
851 multispecies coalescent. Trends Ecol. Evol. 24:332–340.
852 Di Vincenzo V., Gruenstaeudl M., Nauheimer L., Wondafrash M., Kamau P., Demissew S.,
853 Borsch T. 2018. Evolutionary diversification of the African achyranthoid clade
854 (Amaranthaceae) in the context of sterile flower evolution and epizoochory. Ann. Bot.
855 122:69–85.
856 Dohm J.C., Minoche A.E., Holtgräwe D., Capella-Gutiérrez S., Zakrzewski F., Tafer H., Rupp
857 O., Sörensen T.R., Stracke R., Reinhardt R., Goesmann A., Kraft T., Schulz B., Stadler
858 P.F., Schmidt T., Gabaldón T., Lehrach H., Weisshaar B., Himmelbauer H. 2014. The
859 genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature.
860 505:546–549.
861 Doyle J.J. 1992. Gene Trees and Species Trees: Molecular Systematics as One-Character
862 Taxonomy. Syst. Bot. 17:144.
863 Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between
864 Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.
865 Edwards S.V. 2009. Is A New and General Theory of Molecular Systematics Emerging?
866 Evolution. 63:1–19.
867 Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S.,
868 Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and
869 testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol.
870 Phylogenet. Evol. 94:447–462.

871 Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L.K. 2018. DGEN: A Test Statistic
872 for Detection of General Introgression Scenarios. WABI.
873 Elworth R.A.L., Ogilvie H.A., Zhu J., Nakhleh L. 2019. Advances in Computational Methods for
874 Phylogenetic Networks in the Presence of Hybridization. In: Warnow T., editor.
875 Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret. Cham:
876 Springer International Publishing. p. 317–360.
877 Erfan Sayyari, Siavash Mirarab. 2018. Testing for Polytomies in Phylogenetic Species Trees
878 Using Quartet Frequencies. Genes. 9:132.
879 Flowers T.J., Colmer T.D. 2015. Plant salt tolerance: adaptations in halophytes. Ann. Bot.
880 115:327–331.
881 Folk R.A., Mandel J.R., Freudenstein J.V. 2017. Ancestral Gene Flow and Parallel Organellar
882 Genome Capture Result in Extreme Phylogenomic Discord in a Lineage of Angiosperms.
883 Syst. Biol. 66:320-337.
884 Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X.,
885 Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R.,
886 Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015.
887 Extensive introgression in a malaria vector species complex revealed by phylogenomics.
888 Science. 347:1258524.
889 Foster P.G. 2004. Modeling Compositional Heterogeneity. Syst. Biol. 53:485–495.
890 Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans.
891 R. Soc. B Biol. Sci. 363:4023–4029.

892 Gitzendanner M.A., Soltis P.S., Yi T.-S., Li D.-Z., Soltis D.E. 2018. Plastome Phylogenetics: 30
893 Years of Inferences Into Plant Evolution. Plastid Genome Evolution. Elsevier. p. 293–
894 313.
895 Gonçalves D.J.P., Simpson B.B., Ortiz E.M., Shimizu G.H., Jansen R.K. 2019. Incongruence
896 between gene trees and species trees and phylogenetic signal variation in plastid genes.
897 Mol. Phylogenet. Evol. 138:219–232.
898 Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan
899 L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N.,
900 di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011.
901 Full-length transcriptome assembly from RNA-Seq data without a reference genome.
902 Nat. Biotechnol. 29:644–652.
903 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
904 W., Fritz M.H.Y., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-
905 Bonet T., Alkan C., Prufer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-
906 Petri A., Butthof A., Hober B., Hoffner B., Siegemund M., Weihmann A., Nusbaum C.,
907 Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic
908 D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla
909 M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney
910 E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Paabo S.
911 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
912 Hejase H.A., Liu K.J. 2016. A scalability study of phylogenetic network inference methods using
913 empirical datasets and simulations involving a single reticulation. BMC Bioinformatics.
914 17:422.

915 Hejase H.A., VandePol N., Bonito G.M., Liu K.J. 2018. FastNet: Fast and Accurate Statistical
916 Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data. Comp.
917 Genomics.:242–259.
918 Hernández-Ledesma P., Berendsohn W.G., Borsch T., Mering S.V., Akhani H., Arias S.,
919 Castañeda-Noa I., Eggli U., Eriksson R., Flores-Olvera H., Fuentes-Bazán S., Kadereit
920 G., Klak C., Korotkova N., Nyffeler R., Ocampo G., Ochoterena H., Oxelman B.,
921 Rabeler R.K., Sanchez A., Schlumpberger B.O., Uotila P. 2015. A taxonomic backbone
922 for the global synthesis of species diversity in the angiosperm order Caryophyllales.
923 Willdenowia. 45:281.
924 Hoang D.T., Chernomor O. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation.
925 Mol. Biol. Evol. 35:518–522.
926 Hohmann S., Kadereit J.W., Kadereit G. 2006. Understanding Mediterranean-Californian
927 disjunctions: molecular evidence from Chenopodiaceae-Betoideae. TAXON. 55:67–78.
928 Holder M.T., Anderson J.A., Holloway A.K. 2001. Difficulties in Detecting Hybridization. Syst.
929 Biol. 50:978–982.
930 Huang H., Knowles L.L. 2009. What Is the Danger of the Anomaly Zone for Empirical
931 Phylogenetics? Syst. Biol. 58:527–536.
932 Huang W., Zhou G., Marchand M., Ash J.R., Morris D., Van Dooren P., Brown J.M., Gallivan
933 K.A., Wilgenbusch J.C. 2016. TreeScaper: Visualizing and Extracting Phylogenetic
934 Signal from Sets of Trees. Mol. Biol. Evol. 33:3314–3316.
935 Jarvis D.E., Ho Y.S., Lightfoot D.J., Schmöckel S.M., Li B., Borm T.J.A., Ohyanagi H., Mineta
936 K., Michell C.T., Saber N., Kharbatia N.M., Rupper R.R., Sharp A.R., Dally N.,
937 Boughton B.A., Woo Y.H., Gao G., Schijlen E.G.W.M., Guo X., Momin A.A., Negrão

S., Al-Babili S., Gehring C., Roessner U., Jung C., Murphy K., Arold S.T., Gojobori T., Linden C.G.V.D., van Loo E.N., Jellen E.N., Maughan P.J., Tester M. 2017. The genome of *Chenopodium quinoa*. Nature. 542:307–312.

Kadereit G., Ackerly D., Pirie M.D. 2012. A broader model for C_4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). Proc. R. Soc. B Biol. Sci. 279:3304–3311.

Kadereit G., Borsch T., Weising K., Freitag H. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C_4 Photosynthesis. Int. J. Plant Sci. 164:959–986.

Kadereit G., Hohmann S., Kadereit J.W. 2006. A synopsis of Chenopodiaceae subfam. Betoideae and notes on the taxonomy of *Beta*. Willdenowia. 36:9–19.

Kadereit G., Newton R.J., Vandelook F. 2017. Evolutionary ecology of fast seed germination—A case study in Amaranthaceae/Chenopodiaceae. Perspect. Plant Ecol. Evol. Syst. 29:1–11.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

Kamneva O.K., Rosenberg N.A. 2017. Simulation-Based Evaluation of Hybridization Network Reconstruction Methods in the Presence of Incomplete Lineage Sorting. Evol. Bioinforma. 13:117693431769193.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012.

961 Geneious Basic: An integrated and extendable desktop software platform for the
962 organization and analysis of sequence data. Bioinformatics. 28:1647–1649.
963 Knowles L.L., Huang H., Sukumaran J., Smith S.A. 2018. A matter of phylogenetic scale:
964 Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene
965 tree discord in recent versus deep diversification histories. Am. J. Bot. 105:376–384.
966 Koenen, E., Ojeda, D., Steeves, R., Migliore, J., Bakker, F., Wieringa, J., Kidner, C., Hardy, O.,
967 Pennington, R., Bruneau, A., Hughes, C. 2020. Large-scale genomic sequence data
968 resolve the deepest divergences in the legume phylogeny and support a near-
969 simultaneous evolutionary origin of all six subfamilies. New Phytol. 225: 1355-1369.
970 Kubatko L.S., Chifman J. 2019. An invariants-based method for efficient identification of hybrid
971 species from large-scale genomic data. BMC Evol. Biol. 19:112.
972 Kumar, S., Filipski, A., Battistuzzi, F., Kosakovsky Pond, S., Tamura, K., 2012. Statistics and
973 truth in phylogenomics. Mol. Biol. Evol. 29:457–472.
974 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined Selection of
975 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol.
976 Evol. 29:1695–1701.
977 Lee-Yaw J.A., Grassa C.J., Joly S., Andrew R.L., Rieseberg L.H. 2019. An evaluation of
978 alternative explanations for widespread cytonuclear discordance in annual sunflowers
979 (*Helianthus*). New Phytol. 221:515–526.
980 Li, B., Wang, J., Yao, L., Meng, Y., Ma, X., Si, E., Ren, P., Yang, K., Shang, X., Wang, H.
981 Halophyte *Halogeton glomeratus*, a promising candidate for phytoremediation of heavy
982 metal-contaminated saline soils. Plant Soil. 442:323–331.

983 Lightfoot D.J., Jarvis D.E., Ramaraj T., Lee R., Jellen E.N., Maughan P.J. 2017. Single-molecule
984 sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus*
985 *hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol.
986 15:74.
987 Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the Anomaly Zone in Species Trees
988 and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny (Squamata:
989 Scincidae). Syst. Biol. 65:465–477.
990 Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics. 26:962–
991 963.
992 Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial Phylogenomics of Early Land
993 Plants: Mitigating the Effects of Saturation, Compositional Heterogeneity, and Codon-
994 Usage Bias. Syst. Biol. 63:862–878.
995 Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:532–536.
996 Masson R., Kadereit G. 2013. Phylogeny of Polycnemoideae (Amaranthaceae): Implications for
997 biogeography, character evolution and taxonomy. TAXON. 62:100–111.
998 Maureira-Butler I.J., Pfeil B.E., Muangprom A., Osborn T.C., Doyle J.J. 2008. The Reticulate
999 History of *Medicago* (Fabaceae). Syst. Biol. 57:466–482.
1000 Mclean B.S., Bell K.C., Allen J.M., Helgen K.M., Cook J.A. 2019. Impacts of Inference Method
1001 and Data set Filtering on Phylogenomic Resolution in a Rapid Radiation of Ground
1002 Squirrels (Xerinae: Marmotini). Syst. Biol. 68:298–316.
1003 Meyer B.S., Matschiner M., Salzburger W. 2017. Disentangling Incomplete Lineage Sorting and
1004 Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. Syst.
1005 Biol. 66:531–550.

1006 Minh B.Q., Hahn M., Lanfear R. 2020. New methods to calculate concordance factors for
1007 phylogenomic datasets. Mol. Biol. Evol. msaa106
1008 Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating Summary Methods for Multilocus
1009 Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol.
1010 65:366–380.
1011 Morales-Briones D.F., Liston A., Tank D.C. 2018a. Phylogenomic analyses reveal a deep history
1012 of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New
1013 Phytol. 218:1668–1684.
1014 Morales-Briones D.F., Romoleroux K., Kolář F., Tank D.C. 2018b. Phylogeny and Evolution of
1015 the Neotropical Radiation of *Lachemilla* (Rosaceae): Uncovering a History of Reticulate
1016 Evolution and Implications for Infrageneric Classification. Syst. Bot. 43:17–34.
1017 Moray C., Goolsby E.W., Bromham L. 2016. The Phylogenetic Association Between Salt
1018 Tolerance and Heavy Metal Hyperaccumulation in Angiosperms. Evol. Biol. 43:119–
1019 130.
1020 Mower J.P. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes,
1021 chloroplast genes and user-defined alignments. Nucleic Acids Res. 37:W253–W259.
1022 Müller K., Borsch T. 2005. Phylogenetics of Amaranthaceae Based on *matK/trnK* Sequence
1023 Data: Evidence from Parsimony, Likelihood, and Bayesian Analyses. Ann. Mo. Bot.
1024 Gard. 92:66–102.
1025 Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: A Fast and Effective
1026 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol.
1027 32:268–274.

1028 Osuna-Mascaró C., Rubio de Casas R., Perfectti F. 2018. Comparative assessment shows the
1029 reliability of chloroplast genome assembly using RNA-seq. Sci. Rep. 8:17404.
1030 Pamilo P., Nei M. 1988. Relationships between Gene Trees and Species Trees. Mol. Biol. Evol.
1031 5:568–583.
1032 Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet Sampling
1033 distinguishes lack of support from conflicting support in the green plant tree of life. Am.
1034 J. Bot. 105:385–403.
1035 Piirainen M., Liebisch O., Kadereit G. 2017. Phylogeny, biogeography, systematics and
1036 taxonomy of Salicornioideae (Amaranthaceae/Chenopodiaceae) – A cosmopolitan, highly
1037 specialized hygrohalophyte lineage dating back to the Oligocene. Taxon. 66:109–132.
1038 Prasanna A.N., Gerber D., Kijpornyongpan T., Aime M.C., Doyle V.P., Nagy L.G. 2020. Model
1039 Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep
1040 Basidiomycota Relationships. 69:17–37
1041 R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna,
1042 Austria: R Foundation for Statistical Computing.
1043 Rannala B., Yang Z. 2003. Bayes Estimation of Species Divergence Times and Ancestral
1044 Population Sizes Using DNA Sequences From Multiple Loci. Genetics. 166:1645–1656.
1045 Rieseberg L.H., Soltis D.E. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants.
1046 Evol. Trends Plants. 5:65–84.
1047 Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
1048 Rosenberg N.A. 2013. Discordance of Species Trees with Their Most Likely Gene Trees: A
1049 Unifying Principle. Mol. Biol. Evol. 30:2709–2713.

1050 Roycroft E.J., Moussalli A., Rowe K.C. 2020. Phylogenomics Uncovers Confidence and
1051 Conflict in the Rapid Radiation of Australo-Papuan Rodents. Syst. Biol. 69:431–444.
1052 Salichos L., Stamatakis A., Rokas A. 2014. Novel Information Theory-Based Measures for
1053 Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31:1261–1271.
1054 Sang T., Crawford D.J., Stuessy T.F. 1995. Documentation of reticulate evolution in peonies
1055 (Paeonia) using internal transcribed spacer sequences of nuclear ribosomal DNA:
1056 implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. 92:6813–
1057 6817.
1058 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
1059 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
1060 Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.
1061 Schliesky S., Gowik U., Weber A.P.M., Bräutigam A. 2012. RNA-Seq Assembly – Are We
1062 There Yet? Front. Plant Sci. 3:220.
1063 Schwarz G. 1978. Estimating the Dimension of a Model. Ann. Stat. 6:461–464.
1064 Sharp P.M., Li W.-H. 1986. An evolutionary perspective on synonymous codon usage in
1065 unicellular organisms. J. Mol. Evol. 24:28–38.
1066 Shi C., Wang S., Xia E.-H., Jiang J.-J., Zeng F.-C., Gao L.-Z. 2016. Full transcription of the
1067 chloroplast genome in photosynthetic eukaryotes. Sci. Rep. 6:30135.
1068 Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst.
1069 Biol. 51:492–508.
1070 Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree
1071 selection. Bioinformatics. 17:1246–1247.

1072 Smith D.R. 2013. RNA-Seq data: a goldmine for organelle research. Brief. Funct. Genomics.
1073 12:454–456.
1074 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals
1075 conflict, concordance, and gene duplications with examples from animals and plants.
1076 BMC Evol. Biol. 15:745.
1077 Smith S.A., O'Meara B.C. 2012. treePL: divergence time estimation using penalized likelihood
1078 for large phylogenies. Bioinformatics. 28:2689–2690.
1079 Solís-Lemus C., Ané C. 2016a. Inferring Phylogenetic Networks with Maximum
1080 Pseudolikelihood under Incomplete Lineage Sorting. PLOS Genet. 12:e1005896.
1081 Soltis D.E., Kuzoff R.K. 1995. Discordance between nuclear and chloroplast phylogenies in the
1082 Heuchera group (Saxifragaceae). Evolution. 49:727–742.
1083 Srivastava S.K. 1969. Assorted angiosperm pollen from the Edmonton Formation
1084 (Maestrichtian), Alberta, Canada. Can. J. Bot. 47:975–989.
1085 Stamatakis A. 2014. RAxML version 8 - a tool for phylogenetic analysis and post-analysis of
1086 large phylogenies. Bioinformatics. 30:1312–1313.
1087 Sugiura N. 1978. Further analysts of the data by akaike' s information criterion and the finite
1088 corrections. Commun. Stat. - Theory Methods. 7:13–26.
1089 Swofford D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods) version
1090 4. Sunderland MA Sinauer Assoc.
1091 Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and
1092 reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:322–16.
1093 The Angiosperm Phylogeny Group, Chase M.W., Christenhusz M.J.M., Fay M.F., Byng J.W.,
1094 Judd W.S., Soltis D.E., Mabberley D.J., Sennikov A.N., Soltis P.S., Stevens P.F. 2016.

1095 An update of the Angiosperm Phylogeny Group classification for the orders and families
1096 of flowering plants: APG IV. Bot. J. Linn. Soc. 181:1–20.
1097 Vargas O.M., Ortiz E.M., Simpson B.B. 2017. Conflicting phylogenomic signals reveal a pattern
1098 of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae:
1099 *Diplostephium*). New Phytol. 214:1736–1750.
1100 Walker J.F., Walker-Hale N., Vargas O.M., Larson D.A., Stull G.W. 2019. Characterizing gene
1101 tree conflict in plastome-inferred phylogenies. PeerJ. 7:e7747.
1102 Walker J.F., Yang Y., Feng T., Timoneda A., Mikenas J., Hutchison V., Edwards C., Wang N.,
1103 Ahluwalia S., Olivieri J., Walker-Hale N., Majure L.C., Puente R., Kadereit G.,
1104 Lauterbach M., Eggli U., Flores-Olvera H., Ochoterena H., Brockington S.F., Moore
1105 M.J., Smith S.A. 2018. From cacti to carnivores: Improved phylotranscriptomic sampling
1106 and hierarchical homology inference provide further insight into the evolution of
1107 Caryophyllales. Am. J. Bot. 105:446–462.
1108 Wen D., Yu Y., Hahn M.W., Nakhleh L. 2016. Reticulate evolutionary history and extensive
1109 introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol.
1110 25:2361–2372.
1111 Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet.
1112 Syst. Biol. 67:735–740.
1113 Widhelm T.J., Grewe F., Huang J.-P., Mercado-Díaz J.A., Goffinet B., Lücking R., Moncada B.,
1114 Mason-Gamer R., Lumbsch H.T. 2019. Multiple historical processes obscure
1115 phylogenetic relationships in a taxonomically difficult group (Lobariaceae, Ascomycota).
1116 Sci. Rep. 9:8968.

1117 Xu B., Yang Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent
1118 Model. Genetics. 204:1353–1368.
1119 Xu C., Jiao C., Sun H., Cai X., Wang X., Ge C., Zheng Y., Liu W., Sun X., Xu Y., Deng J.,
1120 Zhang Z., Huang S., Dai S., Mou B., Wang Q., Fei Z., Wang Q. 2017. Draft genome of
1121 spinach and transcriptome diversity of 120 Spinacia accessions. Nat. Commun. 8:15275.
1122 Yang Y., Moore M.J., Brockington S.F., Timoneda A., Feng T., Marx H.E., Walker J.F., Smith
1123 S.A. 2017. An Efficient Field and Laboratory Workflow for Plant Phylotranscriptomic
1124 Projects. Appl. Plant Sci. 5:1600128.
1125 Yang Y., Smith S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes
1126 and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for
1127 Phylogenomics. Mol. Biol. Evol. 31:3081–3092.
1128 Yao G., Jin J.-J., Li H.-T., Yang J.-B., Mandala V.S., Croley M., Mostow R., Douglas N.A.,
1129 Chase M.W., Christenhusz M.J.M., Soltis D.E., Soltis P.S., Smith S.A., Brockington S.F.,
1130 Moore M.J., Yi T.-S., Li D.-Z. 2019. Plastid phylogenomic insights into the evolution of
1131 Caryophyllales. Mol. Phylogenet. Evol. 134:74–86.
1132 Yu Y., Degnan J.H., Nakhleh L. 2012. The Probability of a Gene Tree Topology within a
1133 Phylogenetic Network with Applications to Hybridization Detection. PLoS Genet.
1134 8:e1002660–10.
1135 Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate
1136 evolutionary histories. Proc. Natl. Acad. Sci. 111:16448–16453.
1137 Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks.
1138 BMC Genomics. 16:S10.

1139 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
1140 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:523.
1141 Zhao T., Schranz M.E. 2019. Network-based microsynteny analysis identifies major differences
1142 and genomic outliers in mammalian and angiosperm genomes. Proc. Natl. Acad. Sci.
1143 116:2165–2174.
1144 Zheng Y., Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide
1145 parameter space. BMC Bioinformatics. 19:10.
1146 Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., Looz, M. V., Rokas, A. 2020. Quartet-based
1147 computations of internode certainty provide robust measures of phylogenetic
1148 incongruence. Syst. Biol. 69:308–324.
1149 Zhu J., Liu X., Ogilvie H.A., Nakhleh L.K. 2019. A divide-and-conquer method for scalable
1150 phylogenetic network inference from multilocus data. Bioinformatics. 35:i370–i378.